Analyzing/Forecasting Local Elections in Turkey

Mert Emin Kalender, 20702164

December 17, 2012

Outline

- Introduction
  - Project Understanding: Local Elections in Turkey
- Data Understanding
  - 2009 Local Election Results
- Data Preparation
- Modeling
- Evaluation
- Deployment / Demo
- Conclusion
- References

Introduction

- Analysis of people’s political choices
  - **popular** research area in political science
  - reasoning behind world views of individuals
  - voting, religious, consumer behavior etc.
- Turkish voters’ political choices
  - effect of personal values, political campaign etc.
  - after 2003, more analysis of election results
  - mostly about general elections
- This project focuses on **local elections**

Introduction

Turkish Local Elections

- Elections to choose local officials
  - including metropoles, cities, various sized towns
  - 16 metropole, 65 city, 957 town, 1,974 small-sized town
- Hold every 5 years
  - recently some statistics of previous elections are published
- Last one: March 29, 2009
  - ~3,000 officials selected
- Next one: March 30, 2014

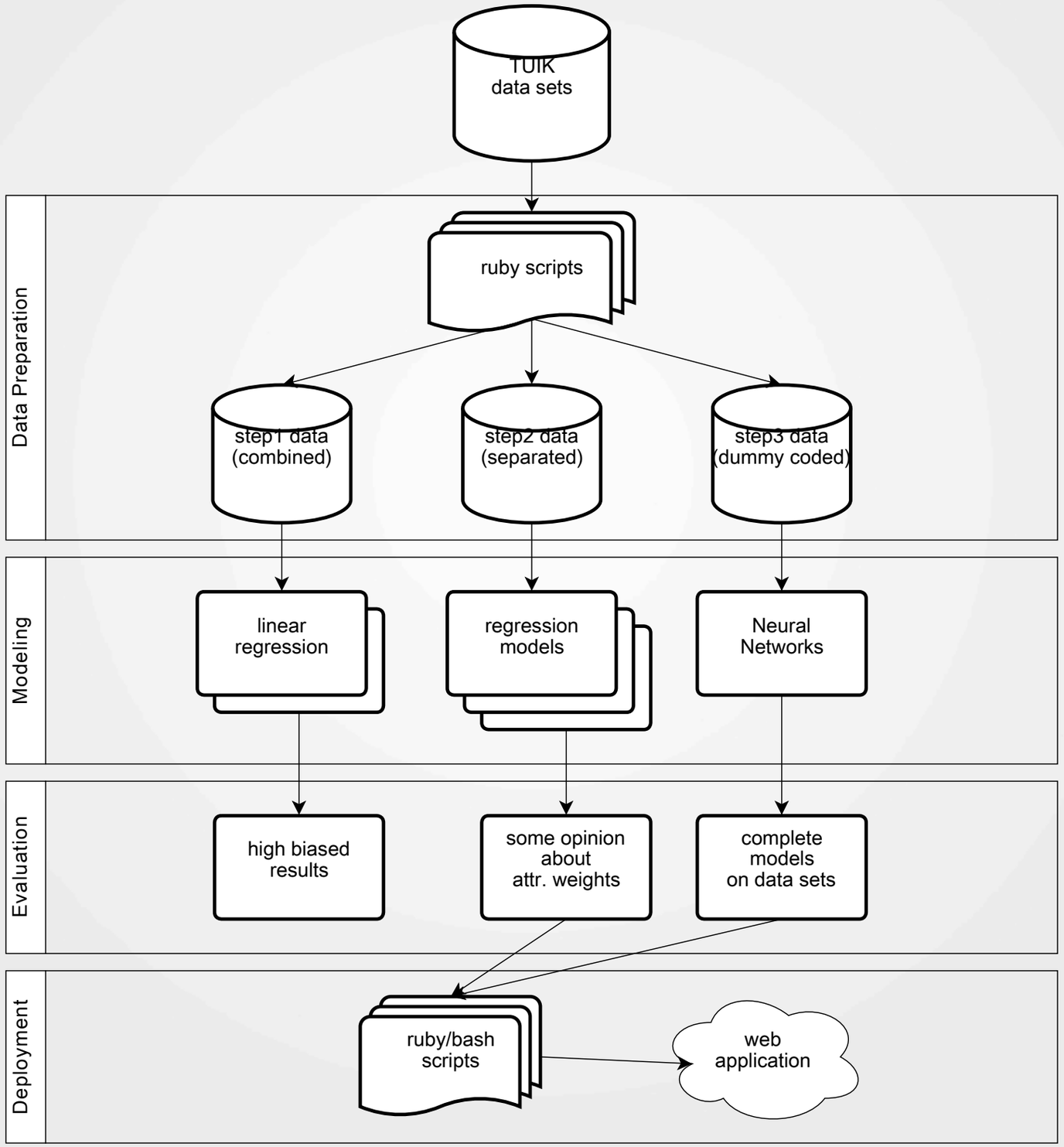
Introduction

Project Understanding

- Domain: **Politics**
- Benefit: An detailed analysis of election results
  - for each city based on following **candidate attributes**:
    age, educational background, marital status and political party
- Solution: A system **predicting chance of winning**
  - for a candidate given age, educational, marital, political party information
- **Regression** problem
  - value of interest is numerical

Introduction

Project Understanding



Data Understanding

2009 Election Results

- Turkish Statistical Institute (TUIK) published the data
- **A data set for each city**
  - with two tables: candidate statistics and winner statistics
  - **consisting of candidates for all local administrations of the city**
- Data is provided in **different formats**
  - HTML, pdf and xls
  - **HTML** is used and parsed with scripts
- Example data set (for Ankara)
  - age, educational background, marital status, political party

Data Understanding

2009 Election Results

- Syntactically and semantically **accurate** data
- No missing values
- No complex data types
- No need for dimensionality reduction
- No need for feature selection
- **Clean** and **simple** data sets
  - **Is this enough?**

Data Preparation

Step 1

- Given all candidate and winner statistics combined
- In total, 4 tables* are created
  - age, educational background, marital status, political party

Age	Sex	City	Voter (%)	Candidate	Winner (%)
25-29	F	C24	6.033	0.0	0.0
30-34	M	C24	5.634	10.87	0.0
30-34	F	C24	5.145	0.0	0.0
35-39	M	C24	5.398	9.42	3.448
35-39	F	C24	5.265	0.725	0.0
40-44	M	C24	4.634	13.768	17.241
40-44	F	C24	4.483	0.725	0.0
45-49	M	C24	4.089	18.116	31.034
45-49	F	C24	4.127	0.725	0.0
50-54	M	C24	3.616	16.667	20.69
50-54	F	C24	3.561	0.725	0.0
55-59	M	C24	3.01	13.768	20.69
55-59	F	C24	2.936	0.0	0.0
60-64	M	C24	2.362	4.348	3.448

Party	Sex	City	LocalR (%)	Candidate	Winner (%)
AKP	M	C11	33.558	17.857	73.333
AKP	F	C11	33.558	0.0	0.0
DP	M	C46	4.139	5.97	8.065
DP	F	C46	4.139	0.0	0.0
CHP	M	C46	6.846	8.955	11.29
CHP	F	C46	6.846	0.0	0.0
DSP	M	C46	1.574	5.672	4.839
DSP	F	C46	1.574	0.299	0.0
SP	M	C46	3.644	8.955	3.226
SP	F	C46	3.644	0.0	0.0
MHP	M	C46	24.851	17.612	11.29
MHP	F	C46	24.851	0.299	0.0
BGZ	M	C46	0.062	0.896	0.0
BGZ	F	C46	0.062	0.0	0.0

* Data tables in reality are bigger than the listed ones. They are minimized for the sake of simplicity.

Modeling

Step 1

- Various **regression** models applied [rm]
  - linear regression result for **age data set** of Ankara

```
0.186 * Age
+ 1.553 * Sex
+ 1.246 * Candidate
- 2.824
```

  - predicts a win value between 4%-8% (with squared error > 50%)
  - linear regression result for **educational background data set** of Ankara

```
- 3.016 * Education
+ 2.950 * Sex
+ 1.172 * Candidate
+ 0.896
```

  - predicts a win value between 12%-19% (with squared error > 55%)
  - **similar results for other data sets as well**

Evaluation

Step 1

- **Biased models** obtained
  - Not enough data
  - Polynomial city information is a problem
    - city attribute expressed as `c{#plateNumber}`
  - Voter/LocalR features not really acceptable
    - values for the main city, (but each local place has its own)
    - removed with the expert consultation
- **No good results!**

Data Preparation

Step 2

- Data for each city separated from the data sets
- Data generation performed
- Each city with 4 tables* for
  - age, educational background, marital status, political party

Age	Sex	Win
25-29	M	1
30-34	M	1
30-34	M	1
35-39	M	1
40-44	F	0
40-44	M	0
45-49	M	0
45-49	F	1
45-49	M	1
50-54	M	0
50-54	F	1
55-59	F	0
60-64	M	1
75+	M	1

Education	Sex	Win
E3	M	1
E3	M	1
E3	M	1
E3	M	1
E3	F	0
E3	M	0
E2	M	0
E2	F	1
E1	M	1
E1	M	0
E1	F	1
E2	F	0
E3	M	1
E3	F	0

M. Status	Sex	Win
S	M	1
S	M	1
S	M	1
M	M	1
D	F	0
D	M	0
W	M	0
M	F	1
M	M	1
S	M	0
D	F	1
W	F	0
S	M	1
S	F	0

P.Party	Sex	Win
DSP	M	1
DTP	M	1
BTP	M	1
CHP	M	1
AKP	F	0
AKP	M	0
MHP	M	0
MHP	F	1
MHP	M	1
SP	M	0
CHP	F	1
CHP	F	0
CHP	M	1
CHP	F	0

* Data tables in reality are bigger than the listed ones. They are minimized for the sake of simplicity.

Modeling

Step 2

- **Regression** models (re)performed [rm]
  - Categorical values treated as ordinal
  - linear/polynomial regression result for **age data set** of Ankara

```
- 0.140 * Age
+ 0.140
```

```
- 0.000 * Age ^ 2.000
- 0.376 * Sex ^ 2.000
+ 0.168
```

  - linear/polynomial regression result for **political party data set** of Ankara

```
0.013 * Party
- 0.137 * Sex
+ 0.033
```

```
0.013 * Party ^ 1.000
- 0.134 * Sex ^ 4.000
+ 0.033
```

- logistic regression result for **marital status data set** of Ankara

```
Bias (offset): -189.755
w[Marital] = 240.427
w[Sex] = 222.459
```

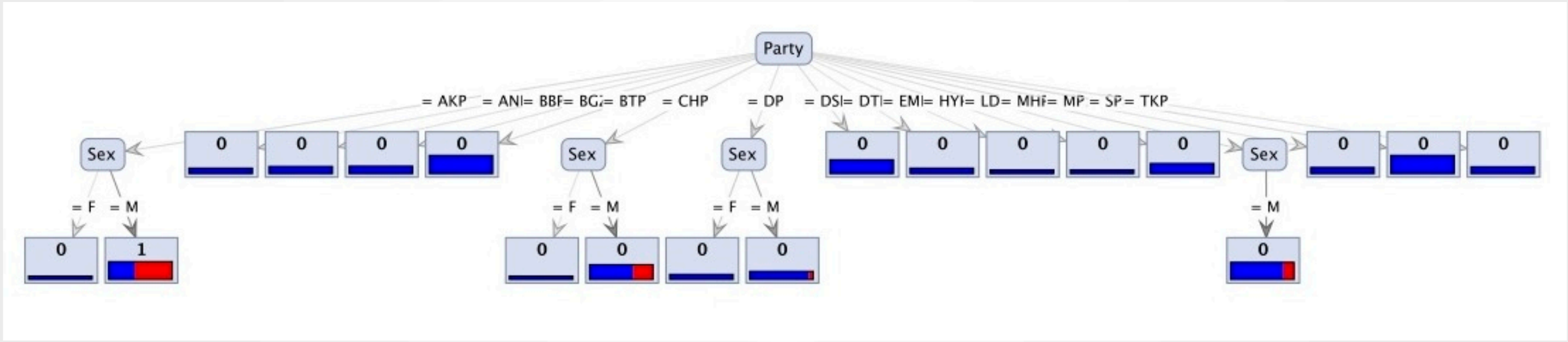
- logistic regression result for **political party data set** of Ankara

```
Bias (offset): -14.888
w[Party] = 420.013
w[Sex] = 37.619
```

Modeling

Step 2

- **ID3 Tree** performed on data sets [rm]
  - result for **political party data set** of Ankara (accuracy: ~90%, precision/recall: ~60%)



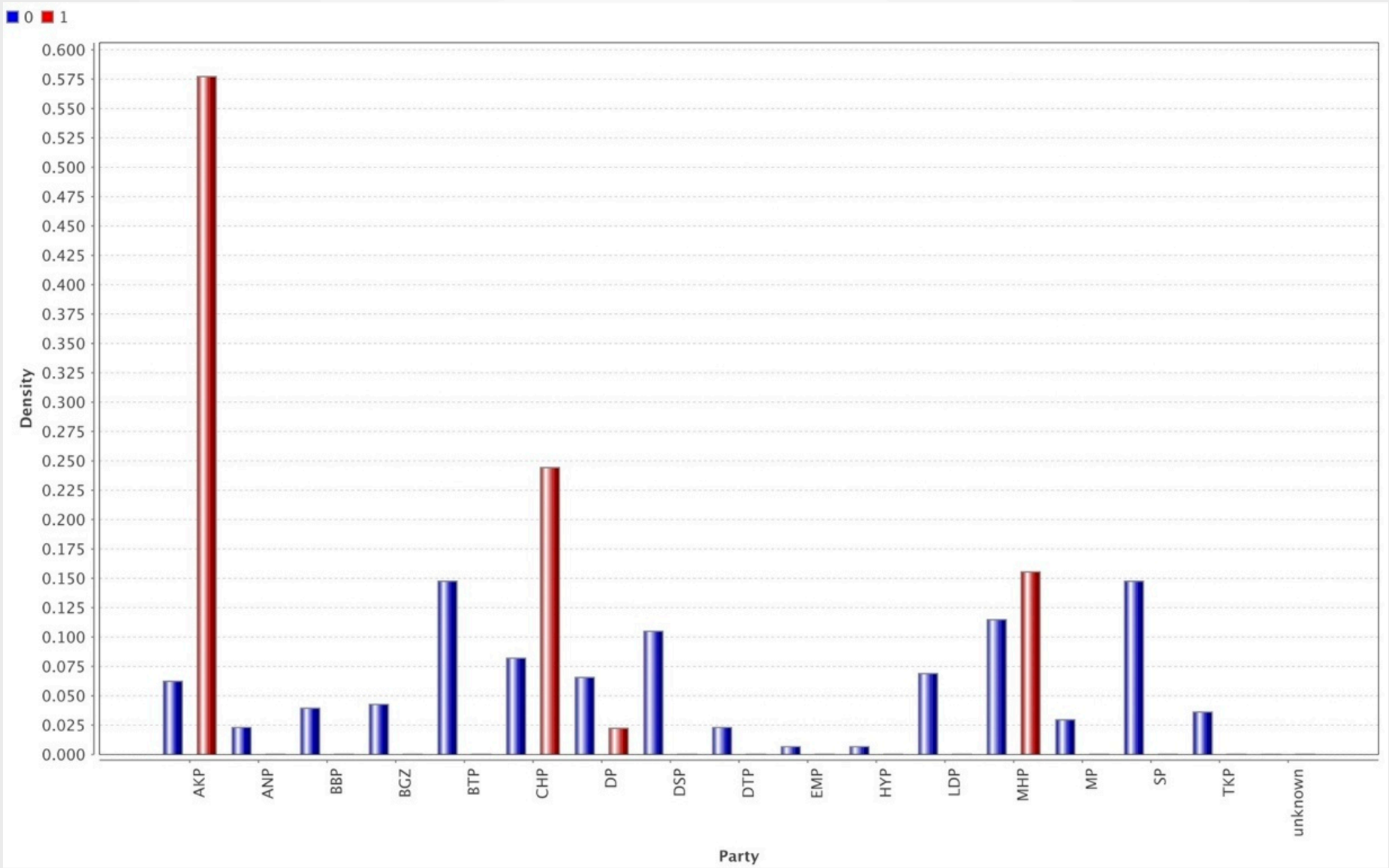

```
Party = AKP
|   Sex = F: 0 {0=1, 1=0}
|   Sex = M: 1 {0=18, 1=26}
Party = CHP
|   Sex = F: 0 {0=1, 1=0}
|   Sex = M: 0 {0=24, 1=11}
Party = DP
|   Sex = F: 0 {0=6, 1=0}
|   Sex = M: 0 {0=14, 1=1}
Party = MHP
|   Sex = M: 0 {0=35, 1=7}
```

```
Party = ANP: 0 {0=7, 1=0}
Party = BBP: 0 {0=12, 1=0}
Party = BGZ: 0 {0=13, 1=0}
Party = BTP: 0 {0=45, 1=0}
Party = DSP: 0 {0=32, 1=0}
Party = DTP: 0 {0=7, 1=0}
Party = EMP: 0 {0=2, 1=0}
Party = HYP: 0 {0=2, 1=0}
Party = LDP: 0 {0=21, 1=0}
Party = MP: 0 {0=9, 1=0}
Party = SP: 0 {0=45, 1=0}
```

Modeling

Step 2

- **Naive Bayes** performed on data sets [rm]
  - result for **political party data set** of Ankara (accuracy: ~90%, precision/recall: ~60%)



Modeling

Step 2

- **RIMARC** applied on data sets [ri]
  - result for **political party data set** of Ankara

```
Party (weight: 0.8378)
If Party="AKP" Then Risk=0.5777778 (45)
If Party="CHP" Then Risk=0.30555555 (36)
If Party="MHP" Then Risk=0.16666667 (42)
If Party="DP" Then Risk=0.04761905 (21)
If Party="BGZ" Then Risk=0.0 (13)
If Party="EMP" Then Risk=0.0 (2)
If Party="TKP" Then Risk=0.0 (11)
If Party="LDP" Then Risk=0.0 (21)
If Party="HYP" Then Risk=0.0 (2)
If Party="SP" Then Risk=0.0 (45)
If Party="BTP" Then Risk=0.0 (45)
If Party="DTP" Then Risk=0.0 (7)
If Party="MP" Then Risk=0.0 (9)
If Party="ANP" Then Risk=0.0 (7)
If Party="BBP" Then Risk=0.0 (12)
If Party="DSP" Then Risk=0.0 (32)
```

  - result for **age data set** of Ankara

```
Age (weight: 0.3236)
If Age="40-44" Then Risk=0.2037037 (54)
If Age="45-49" Then Risk=0.17142858 (70)
If Age="55-59" Then Risk=0.16666667 (42)
If Age="50-54" Then Risk=0.14925373 (67)
If Age="65-69" Then Risk=0.125 (8)
If Age="35-39" Then Risk=0.10256410 (39)
If Age="25-29" Then Risk=0.0 (22)
If Age="30-34" Then Risk=0.0 (27)
If Age="60-64" Then Risk=0.0 (16)
If Age="70-74" Then Risk=0.0 (3)
If Age="75+" Then Risk=0.0 (2)
```

Evaluation

Step 2

- **Regression** still does not provide good results
  - still biased
  - **categorical** types changed into **ordinal** ones
    - some of them makes sense (age, educational background)
    - others do not (marital status, political party)
- **ID3 tree** and **Naive Bayes** provides some idea
  - about which values are significant and improve winning chance
- **RIMARC** results may be useful too
  - early to say something at this phase

Data Preparation

Step 3

- Dummy coding applied for all data sets

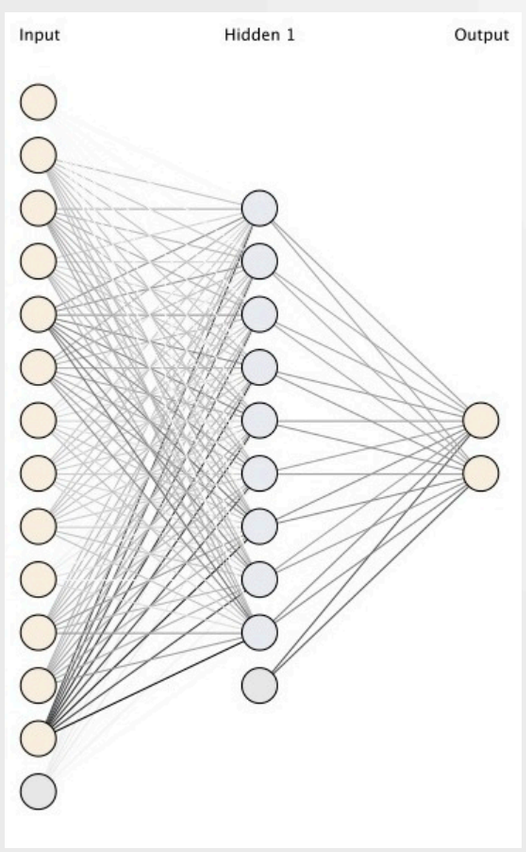
ANP	MHP	LDP	DSP	DTP	IP	CHP	HP	BBP	BDP	MP	ODP	TKP	DP	SP	AKP	BTP	EMP	HYP	BGZ	Sex	Win
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

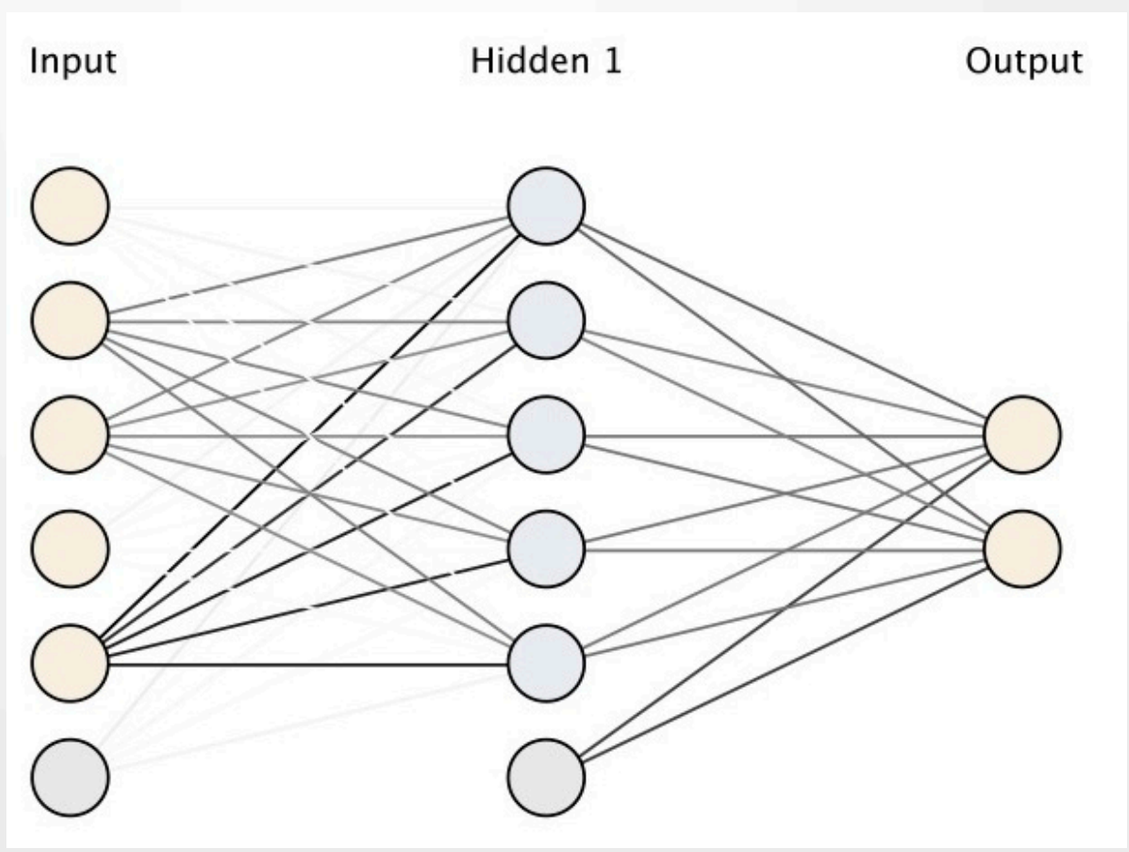
* Data table in reality is bigger than the listed one. It is minimized for the sake of simplicity.

Modeling

Step 3

- **Neural Network** applied [rm]
  - with ten-fold cross validation and stratified sampling
- Four models are generated for each city
  - result for **age/educational background data sets** of Ankara





Evaluation

Step 3

- **Neural Network** provided good set of models
  - able to derive some meaning from the data
  - two output nodes: one for class 0 and one for class 1
  - enables us to say something based on output node values
- **Too few women** candidate/winner
  - 554 among 15,569 candidates
  - 26 among 2,931 winners
- **Another** Neural Network modeling without sex feature

Deployment / Demo

- Requirement of **a way to combine four models**
  - How to combine four of them together and come with a result?
  - **RIMARC** results may be helpful
    - performed again and **weights generated** by the algorithm saved
- A **web application** created
  - **Neural Network** models into ruby classes
  - **RIMARC** results into feature weighting suggestions
  - **statistics** from 2009 into various data graphs
  - [d3, jq, sn, tw] used
- Link to the demo application on [hr]

Conclusion

- From **the data sets of TUIK**
- To highly capable **a web application** on local elections
- Clean and simple data is **not enough**
- **Prepare the data** in a such way that enables good analysis
- Data is **(re)organized** and **(re)generated** several times
- At the end in total (81 * 4) * 2 = **648 NN models** created
- **81 different weight profile** generated via RIMARC
- A good analysis of local elections is now **possible**

References

- [d3]: d3js (Data driven documents)
- [hr]: heroku
- [jq]: jQuery
- [rm]: Rapidminer
- [ri]: RIMARC
- [sn]: sinatra
- [tw]: twitter bootstrap

*

Thanks

* İbrahim Melih Gökçek is a Turkish politician who has been the mayor of Ankara since 1994. He won local elections in 1994, 1999, 2004, and 2009 with three different political parties. (image: http://forum.paticik.com/thumbnails/333/3db/06c/910/8d5/199/930/658/93d/5d1/44_450xNULL.jpg)